FACULTY WORKING PAPER 92-0136

330
B385
1992:136 COPY 2
STX

Maximum Versus Meaningful Discrimination in Scale Responses: A Perspective on the Optimal Number of Response Categories to Use in a Scale

The Library of the
AUG 1 1 1992
University of Illinois

Madhubalan Viswanathan
Department of Business Administration
University of Illinois

Seymour Sudman
Department of Business Administration
University of Illinois

Bureau of Economic and Business Research
College of Commerce and Business Administration
University of Illinois at Urbana-Champaign

BEBR

FACULTY WORKING PAPER NO. 92-0136

College of Commerce and Business Administration

University of Illinois at Urbana-Champaign

June 1992

Maximum Versus Meaningful Discrimination in Scale Response: A Perspective on the Optimal Number of Response Categories to Use in a Scale

Madhubalan Viswanathan
Seymour Sudman

Department of Business Administration

Digitized by the Internet Archive
in 2011 with funding from
University of Illinois Urbana-Champaign

http://www.archive.org/details/maximumversusmea92136visw

Abstract

This paper argues for the use of the number of response categories that are meaningful to respondents as a criterion in designing scales. In contrast to a focus in past research on maximizing the discrimination elicited from respondents, a distinction is drawn in this paper between the number of categories that respondents can reliably discriminate between and the number of categories that respondents typically use in thinking about a continuum (i.e., the meaningful number of categories). The importance of the meaningful number of categories is demonstrated by showing its influence on scale response. Specifically, it is hypothesized that if fewer categories are used by respondents in thinking about a continuum, fewer response categories in a rating scale would be utilized for ratings on this continuum. The first study used a sorting task from research on categorization as an open-ended method to assess the number of categories that individuals typically use in thinking about some product dimensions. A second study collected ratings of these same products along the specific dimensions using scales with different numbers of response categories and found support for the hypothesis. A third study replicated the findings of the second study. The results point to the importance of the meaningful number of categories in designing scales as well as the possibility of extracting such information from scale responses.

Several researchers in the past have addressed the problem of assessing the optimal number of response categories to employ in a scale (cf., Mattell and Jacoby 1972; Lissitz and Green 1975). Cox (1980), in reviewing the literature in this area of research, points out that suggestions made by researchers range from the use of as few as two to as many as 25 alternatives. Approaches in the past include the assessment of psychometric properties of scales with different numbers of response categories, the use of approximately seven response categories based on research on absolute judgments, and the information theoretic approach to determine information transmitted by a scale (Cox 1980). The tradeoff in determining the number of response categories to use in a scale in these approaches has typically been between the information generated by a scale (for purposes of the researcher) and limits on human ability to discriminate beyond a certain number of response categories. In contrast to a focus in past research on maximizing the discrimination elicited from respondents, a distinction is drawn in this paper between the number of categories that respondents can *reliably discriminate* between and the number of response categories that are *meaningful* to respondents.

This paper argues for the use of the number of response categories that are meaningful to respondents as a criterion in designing scales. Specifically, it is hypothesized that the number of categories that respondents use in thinking about certain continua would influence responses to scales along these continua. If fewer categories are used in thinking about a continuum, fewer response categories in a rating scale would be utilized for ratings on this continuum. A first study used an independent, open-ended method (i.e., a sorting task) employed in past research on categorization (cf., Block et al., 1981) to assess the number of categories that are meaningful to respondents on certain dimensions (i.e., the number of categories that respondents typically use in thinking about a continuum). The sorting task assessed the number of categories employed by respondents for sorting products along specified dimensions. Using the results of a sorting task as a basis, a second study assessed responses to ratings of products along these dimensions using scales with different numbers of response categories. Fewer scale points were found to be used for dimensions with fewer meaningful categories, providing evidence that the meaningful number of response categories that respondents use in thinking about a continuum influences the utilization of response categories of scales used to measure ratings on that continuum. A third study replicated the

findings of the second study.

REVIEW AND HYPOTHESIS

Review of Past Research

The literature on the optimal number of response categories to use in a scale has been characterized by several approaches (Cox 1980). The information theoretic approach (cf. Garner and Hake 1951) has related the number of response categories to the amount of information transmitted by a scale. Information transmitted by a scale has been found to increase with an increase in the number of response categories with a leveling off being found in some studies (Cox 1980). The focus of this approach has been on the extent to which respondents utilize response categories to discriminate stimuli since greater utilization of various response categories translates to higher information transmission by a scale. Metric approaches to the optimal number of response categories have assessed the reliability of scales with different number of response categories. Churchill and Peter (1984), based on a meta analysis, found that the reliability of a scale increased with an increase in the number of response categories. However, other researches have not found such a relationship (cf. Bendig, 1953, 1954). In fact, Mattell and Jacoby (1972) have argued for the use of as few as three scale points in a scale. Research on absolute judgments has focused on perceptual judgments along dimensions such as loudness (Miller 1956). This research has been cited as providing support for the use of approximately seven response categories due to a limit on human ability to discriminate beyond seven levels. However, such a generalization has been questioned since this research related to perception rather than memory and used absolute judgments (Cox 1980).

Past research has been characterized by an emphasis on the use of response categories that would lead to the *maximum level of reliable discriminations* by respondents. The implicit or explicit tradeoff in each of these approaches has been in terms of information generated by a scale (for purposes of the researcher) and limits of human ability to discriminate beyond a certain number of response categories. Researchers point out that a scale with too few categories does not allow respondents to use their ability to discriminate while a scale with too many categories may be beyond the respondents' ability to discriminate (cf., Komorita and Graham, 1965). Research on absolute judgments has been cited in support of this constraint in respondents' ability to discriminate. This

constraint is reflected in limits to increases in reliability with an increase in the number of response categories in a scale in metric approaches, and by the limits to the information generated by a scale in information theoretic approaches. Hence, the focus has largely been on eliciting the maximum level of reliable discrimination from respondents within the constraints in human ability to discriminate beyond a certain number of response categories. While a general rule suggesting the use of a certain number of response categories (i.e., 7 plus or minus 2 categories) has been suggested, researchers have pointed out that the optimal number of response categories could vary anywhere from two or three (Jacoby and Mattell, 1971) to 25 (Guilford, 1954; Champney and Marshall, 1939). It appears that the optimal number of response categories could vary as a function of factors such as the type of dimension being measured (for example, the importance of a dimension) and the nature of the respondents (for example, Alba and Hutchinson (1987) suggest that experts may be more discriminating than novices). Hence, methods need to be developed to assess the appropriate number of response categories to use in a scale before scale development. Cox (1980) suggests that there is an immediate need to develop methods at the pretesting stage to evaluate the nature of information being collected using scales with different number of response categories. This is argued to be the case particularly for stimulus-centered scales since response centered scales involve the use of multiple items which increases the effective redundancy of information and the effective variance of a scale (Cox 1980) .

Maximum versus Meaningful Discrimination in Scale Response

Past research is characterized by an emphasis on the maximum level to which respondents can reliably discriminate between stimuli. However, a distinction can be drawn between the *maximum level to which respondents can discriminate* along some continuum and the *meaningful level of discrimination* along the same continuum. The meaningful level of discrimination is used here to refer to the number of categories that individuals typically use in thinking about a continuum (in situations such as those involving the use of information on a product attribute in making a choice or judgment). This distinction is important in that meaningful rather than maximum discriminations made by respondents may predict outcomes of interest such as choice and purchase behavior since maximum discriminations may merely be an artifact of the measurement scale. In using responses to rating scales to understand the level to which respondents discriminate between brands on specific

attributes, response scales with a large number of response categories may elicit a higher level of discrimination than would typically be employed by respondents in tasks involving ratings on a dimension. Researchers have pointed out the informative function served by response scales and argued for the use of open-ended techniques in certain situations such as in scales measuring behavioral frequencies (cf. Sudman and Bradburn, 1982). In a similar vein, more discriminating scales (i.e., scales with more response categories) may elicit finer discriminations across stimuli.

Consider an example of a rating scale used to rate brands on some attribute (such as 'calorie content' of brands of breakfast cereal). Of importance here is usually the rated differences between brands that are meaningful to consumers, with a view to predicting outcomes such as choice and purchase behavior or to acquire diagnostic information for product improvement. If consumers consider two levels of a continuum to be meaningful to them (such as 'high' versus 'low' calorie content), they may still be able to reliably discriminate stimuli on a five point scale, given the availability of five points to distribute their responses. But a difference between a four and a five on such a scale may not translate into a meaningful difference to them for purposes of using information on the continuum in making a choice or a judgment. The inference made from a five point scale would be that a difference between a four and five is a meaningful one to the respondent. However, such a difference may merely be an artifact of the bias introduced by the number of response categories of a response scale.[1] Hence, differences on the five point scale may not provide *a valid basis for making inferences about the decision making process of the respondent* . It should be noted here that the problem is not one of reliability but of the validity of a scale in terms of the underlying construct being measured (i.e., the construct being either meaningful differences or maximum differences within the constraints of human ability). While a scale that elicits the maximum level of discrimination from a respondent may be more reliable than one that has fewer response categories, the latter may be more valid in measuring ratings that are meaningful to respondents, though less reliable (due to less variance associated with fewer response categories).

Implicit in this distinction between meaningful and maximum levels of discrimination are two assumptions. First, it is assumed that respondents naturally use use a stable number of categories in thinking about a specific continuum, i.e., the meaningful number of categories. Second, it is assumed that meaningful discrimination is not the same as maximum discrimination and that it is

possible to discriminate *reliably* to a greater degree than is meaningful to respondents. With regard to the first premise, recent research has related different types of attributes (i.e., concrete versus abstract attributes) to the use of featural representations (i.e., the use of two levels or categories to describe the attribute) versus dimensional representations (i.e., the use of multiple categories to describe an attribute) (Johnson and Fornell 1987). Hence, the authors suggest a relationship between the type of attribute (i.e., concrete versus abstract) and the number of categories used to represent them in memory. In a similar vein, Tversky (1977) suggests that conceptual stimuli are more likely to be represented as features while perceptual stimuli are more likely to be represented as dimensions. Research on the nature of encoding of magnitudes along attributes in memory has suggested the use of a relatively imprecise categories similar to verbal labels to encode information about product attributes presented in a numerical (i.e., relatively, precise) form (Viswanathan and Childers, 1992). Therefore, implicit in these studies is the notion that consumers may use a relatively stable set of categories in thinking about product attributes. The premise here does not preclude the possibility that the set of categories used by respondents may evolve over time but that, unlike scale responses which are based on the relatively spontaneous formation of a rating in response to an item and its response format (i.e.,the scale), the typical number of categories that respondents use in thinking about specific continua may be relatively well formed and stable.

In support of the second premise that meaningful discrimination may be different from maximum discrimination, a large body of evidence has been documented regarding differences in responses obtained as a function of differences in response scales (cf. Poulton, 1989). Parducci (1965) presented the frequency-range model which suggests that responses involve a trade-off between the frequency principle (wherein respondents use each category for a fixed proportion of judgments) and the range principle (wherein respondents use response categories to divide the range which is the difference between the extreme values). Therefore, it appears that small differences in response scales can elicit differing levels of discrimination from respondents. Further, increases in reliability have also been reported with increases in the number of response categories (Churchill and Peter, 1984), suggesting that finer reliable discriminations can be elicited from respondents with an increase in the number of response categories of a scale up to a point. Therefore, in order to dispute this premise, it would have to be suggested that the maximum discrimination that can be made by

respondents is also the most meaningful level of discrimination.

From a cognitive perspective, while maximum discrimination may be a function of the perception of variations in stimuli in the environment, meaningful discrimination may be related to the usefulness of variations in stimuli for purposes such as choice or judgment. In order for maximum discrimination to match meaningful discrimination, it would have to be asserted that the perception of variation among stimuli in the environment translates into a corresponding variation in its meaningfulness. Whereas the perception of variation may lead to a memory that is capable of a certain level of discrimination, such a perception may not influence the meaningfulness attached to this variation, which may be influenced by factors such as the importance of an attribute or the importance of a decision. The research mentioned earlier (i.e., Johnson and Fornell, 1987; Viswanathan and Childers, 1992) provides some indirect evidence for this line of argument. It appears that featural representations may be more likely to be used for concrete attributes in order to conserve processing capacity for large numbers of concrete attributes (Johnson and Fornell, 1987), suggesting that the typical number of categories used in thinking about a continuum may be influenced by concerns about processing capacity. Further, Viswanathan and Childers (1992) argue that relatively precise numerical attribute information may be recoded into a relatively imprecise verbal form to facilitate its use in decision making, with the possibility that the precise numerical information is encoded in memory for making fine discriminations when required. Again, the implication here is that, though imprecise magnitude categories may be used in decision making, consumers may have the ability to make finer discriminations. A similar conclusion can be drawn from research in psychology on comparative judgments (i.e., tasks that require subjects to compare stimuli along some dimension and make a judgment about the magnitudes of stimuli along that dimension) (cf., Banks, 1977). The debate in this area of research has centered around the fine-grained versus coarse-grained nature of representations used to make comparative judgments. Holyoak and Mah (1982) present a model which suggests that working memory contains a coarse-grained scale which is used to make comparative judgments. If the need for a finer discrimination arises, precise information is available in long term memory. To summarize, there appears to be evidence to suggest that consumers may use a relatively stable set of categories to think about specific continua, while having the ability to discriminate to finer extent than is meaningful to them

for tasks such as choice or judgment.

The distinction between maximum discrimination and meaningful discrimination is crucial for stimulus-centred scales. In the case of response centered scales, the aim is to assess the *relative* standings of individuals. The use of multiple items increases the effective redundancy of information (Cox 1980) and the effective variance of the scale. However, for stimulus centered scales, the goal is often to obtain unbiased judgments of stimuli. The imposition of response categories, while potentially leading to reliable discriminations, may elicit responses that do not reflect unbiased judgments that respondents would make in the absence of biasing information regarding the number of response categories to use (in this context, some researchers have argued for the importance of open-ended items to elicit responses (Mattell and Jacoby, 1972)). Such unbiasedness is often important as, for example, where scale responses are used to predict variables such as consumer choice and judgment. For example, ratings may be used to provide diagnostic information on the product attributes on which a particular brand is outperformed by other brands. Therefore, it is important that differences in ratings are meaningful to respondents. While techniques such as regression may be used to relate predictor variables to dependent variables in a correlational sense and are based on the explained variance, for inferences that are not so much based on strength of correlations but on differences between ratings of stimuli, it is important to validly measure meaningful differences along a continuum.

In light of the body of research that documents differences in usage of response categories as a function of factors such as the number of response categories and category descriptors (cf., Poulton, 1989), it is evident that scales can influence responses in many ways. A goal in valid measurement may then be to investigate the nature of magnitude categories employed for various continua as a means to gain insights into differences in the manner in which respondents think about these continua. Such bases can then be used to determine issues in scale development such as the optimal number of response categories to utilize in a scale. The recommendation being argued for here is similar to the suggestion by Cox (1980) that there is an immediate need to develop methods at the pretesting stage to evaluate the nature of information being collected using scales with different number of response categories.

Effect of Meaningful Discrimination on Scale Response

The approach taken here was to first assess the meaningful number of categories used by respondents for some product dimensions with an independent, open-ended method (i.e., a sorting task). The results based on this method provided a set of product dimensions on which the meaningful number of response categories differ. Say, two sets of dimensions with 3 and 4 categories, respectively, were obtained based on the sorting task referred to as 3-category and 4-category dimensions. The issue examined here was the extent to which ratings of products on these two sets of dimensions reflected their differences in terms of the meaningful number of response categories. Such an approach was taken in order to attempt to demonstrate the effect of meaningfulness in scale response, given the lack of past research with such a focus. If differential scale usage is found, then a main effect of the meaningfulness factor in scale response is suggested, providing evidence for the importance of considering meaningful response categories. Further, an implication here is that data from rating scales can be used to assess the effect of meaningfulness. However, if no effect of lesser versus greater dimensions is found, this is suggestive of the lack of an effect of meaningfulness on scale response *and/or* the invalidity of the independent method used here to assess the number of meaningful categories used by respondents.

Considering a scenario where the meaningful number of categories used by respondents to think about a continuum is less than the number of response categories provided in a scale to rate products along that continuum, if scale responses are influenced by the meaningful number of categories, fewer scale points in a rating scale would be used to rate a set of products for the 3-category dimension when compared to the 4-category dimension. Two factors would influence the number of scale points used by respondents, the number of categories in a scale and the meaningful number of categories of the dimension on which ratings are made. The number of response categories in a scale would exert an upward influence whereas the meaningful number of categories would exert a downward influence on the number of response categories that are used by respondents. Past research citing increases in reliability with increases in the number of scale points (cf., Churchill and Peter, 1984) as well as other research (cf., Parducci, 1965) suggest the influence of the number of response categories of a scale on the utilization of these categories by respondents. To the extent that the number of response categories in a scale (say, a 5 point scale) are greater than the meaningful number of categories, fewer response categories may be used for the 3-category

dimension when compared to the 4-category dimension. As the number of scale points are increased, this difference will remain, though the effect of the number of scale points will be to increase the number of response categories used by respondents. (A scenario where the effects are reversed would be when the number of response categories is less than the meaningful number of categories. Therefore, the meaningful number of categories would encourage the use of more response categories but the lack of availability of sufficient categories would force respondents to restrict responses to those available). The discussion to this point leads to the following hypothesis.

H1: Ratings of products on dimensions with fewer meaningful number of categories will involve the usage of fewer scale points in a rating scale when compared to ratings on dimensions with higher meaningful number of categories.

A description of the independent method (i.e., the sorting task) of assessing the meaningful number of categories and the results are reported in Study 1. Study 2 used some product-dimension combinations from Study 1 and scales with different numbers of response categories to assess the prediction suggested above. Study 3 replicated the findings of Study 2.

METHOD

Study 1

Overview.

The focus of this study was to employ an independent method to assess the meaningful number of response categories used by respondents for certain product dimensions. The outcomes based on this method were used as a benchmark to compare responses to rating scales with different numbers of response categories. The independent method used here was derived from research on breadth of categorization in psychology. Bruner and Tajfel (1961) define breadth of category as "the range of stimuli that are placed in the same class or category and share a common label" (p. 231). One type of task involves a set of objects that subjects are required to sort into categories with the specification of a dimension on which the sorting is to be done (Block et al., 1981). The number of categories employed is used as a measure of conceptual differentiation (Gardner and Schoen, 1962) and is argued to be negatively related to the breadth of the categories employed (Block et al., 1981). The categorization of objects on a specified dimension provides a means of understanding the

meaningful number of categories used by respondents to think about a continuum. Extending this notion of the number of categories used to sort objects on a specified dimension to a product attribute continuum, the sorting of products on specified dimensions was used here. This approach was well suited for the purpose at hand since it requires subjects to divide up a specified dimension into several categories *without the imposition of a pre-determined number of response categories.* Therefore, it involved the free elicitation of the number of categories employed to describe an attribute continuum.

Procedure.

In past research using the sorting task, subjects have been instructed to sort objects into groups that "go together" (Block et al. 1981) based on certain dimensions (e.g., Gardner and Schoen (1962). A similar approach was adopted here wherein a particular attribute was specified and subjects were required to sort 12 products/brands belonging to a category into groups. Subjects were presented with a list of the 12 products and asked to write down the names of products and draw circles around them to indicate groups. A small pretest was conducted to assess the effectiveness of the proposed procedure where subjects sorted 12 brands of candy bars, provided written descriptions of how they performed the sorting task, and filled out several scales relating to the sorting task. Written descriptions and response to scales suggested that subjects were adhering to instructions and performing the task with relative ease.

120 students from a midwestern university participated, with three groups of 40 subjects each, completed a questionnaire in Study 1. Each subject performed four sortings for four different product categories (i.e., candy bars, snack foods, soft drinks, and beverages) on the basis of a specified attribute for each category (such as sweetness of soft drinks, caffeine content of beverages, etc.). The product categories, the specific products (i.e., 12 in each category), as well as the attributes used for each product category were based on past research (Johnson et al., 1991). After each sorting task, subjects wrote descriptions of how they performed the task and completed scales relating to the task.

The sorting task was assessed by examining subjects' responses to several scales as well as their written descriptions. The descriptions of the sorting tasks provided by subjects as well as ratings suggested that subjects were adhering to instructions in performing the task based on the

specified attribute, a central requirement to obtain an accurate number of categories. Subjects' confidence, knowledge, experience, motivation, and ease in performing the sorting task were assessed after each sorting and found to be satisfactory. The sorting task was also assessed by treating the number of groups that subjects sorted objects into for each the four sortings as items in a multiple item scale. Moderate inter-correlations between these items (average inter-correlation = 0.46) and a moderate reliability (Coefficient alpha = 0.77) for the four item scale point to the existence of individual differences in sorting which is tapped by the sorting task, a result that is expected in light of past research in psychology using sorting tasks that have demonstrated individual differences in breadth of categorization (cf., Block et al. 1981).

Results.

The mean number of categories employed for each product-dimension combination was computed across subjects. Based on this analysis, several product-dimension combinations appeared to be appropriate for use in subsequent studies. The mean number of sorted groups ranged from 2.75 to 4.20. In particular, four product-dimension combinations were chosen such that two of these could represent 'fewer' number of magnitude categories (referred to as 3-category dimensions) and two could represent 'greater' magnitude categories (referred to as 4-category dimensions). The product-dimension combinations chosen (with dimensions shown in quotations) were 'sweetness' of beverages, 'how good a snack' a snack food is, 'caffeine content' of soft drinks, and 'chocolate flavor' of candy bars, with the mean number of categories employed being 4.13, 4.18, 2.95, and 3.50, respectively, and standard deviations being 1.53, 1.17, 1.00, and 1.18, respectively. Comparisons of 4 versus 3 category dimensions based on the number of sorted groups using t-tests suggested significant differences. Therefore, Study 1 provided a set of product-dimension combinations that appeared to possess different meaningful number of categories and could be used to assess the hypothesis about the effect of meaningfulness on scale response.

Study 2

In Study 2, the product-dimension combinations chosen from Study 1 were used to collect ratings of products on dimensions on scales with different numbers of response categories in order to test the hypothesis that differences in the meaningful number of response categories used for a dimension based on the sorting task would translate into differences in the frequency of utilization of

scale points in a rating task.

Procedure.

150 students at a midwestern university participated in this study with 30 subjects being assigned to each of five groups such that the groups differed in terms of the number of response categories used for the rating scales. 3, 4, 5, 7, and 9 response categories, respectively, were used for the five groups. Subjects rated the set of four product-dimension combinations chosen from Study 1 (i.e., 12 products per dimension for a total of 48 ratings). Scales were end-anchored in terms of the dimension that was being measured (e.g., scales for caffeine content of soft drinks were anchored Very low caffeine content-Very high caffeine content).

Data Analyses.

An important issue in data analysis was to find an appropriate indicator of the frequency of usage of scale points. The frequency of usage of response categories have been aggregated across subjects in past research in order to identify the underutilization of certain categories such as extremes (cf. Wyatt and Meyers, 1987). For example, Wyatt and Meyers (1987) studied Likert scales with different anchors and found less usage for end points that were labeled to convey more "absolute" or extreme valence. The aim of such studies has been to investigate the underutilization or overutilization of specific scale positions for purposes such as comparison of the use of extreme versus moderate verbal anchors. However, the goal here was not to assess usage of particular scale positions but to assess the frequency of usage of response categories *without regard to specific scale positions*. An aggregation of frequency of usage of specific scale positions is inappropriate for the purpose at hand as illustrated by the following example. Considering an extreme scenario where a respondent used a single response category to characterize a range of stimuli, if all respondents were similar in this respect but were equally likely to use any one of the total number of response categories (say, 7), an aggregation across respondents would result in a uniform distribution across response categories suggesting the use of 7 categories. However, with respect to the number of categories utilized by each respondent, the appropriate indicator is one category. By extension, similar problems arise with the use of two or more categories by respondents. Since the goal here was not to identify usage of specific scale positions but to identify the overall number of response categories used in describing a set of stimuli, a different type of analysis was performed.

The frequency of use of each response category was computed for each respondent and these frequencies were ordered from highest to lowest. Each set of ordered frequencies for a respondent reflected the extent to which each response category was used in the order from most to least frequently used. These frequencies were aggregated across subjects since the aggregated values referred to a ranking on frequency of usage such that the most frequently used category was aggregated across subjects followed by the second most frequently used and so on. The number of categories required to capture certain percentages of ratings has been used in past research (Ramsay, 1973). Adapting such an indicator to the purpose at hand, the dependent variable computed using this data was the number of stimuli captured by specific percentages of response categories.[2] By using percentages of response categories rather than the actual number of response categories, a uniform indicator can be used for scales with different numbers of response categories. Considering 3 and 4 category dimensions, the number of stimuli captured by, say, 25% of response categories is hypothesized to be lower for the 4-category dimension when compared to the 3-category dimension. This is because a greater number of response categories are hypothesized to be used for the 4-category dimension, thereby leading to a lower number of stimuli being captured by a certain percentage of response categories.

Results.

The effects of the meaningful number of categories and the number of categories in a scale were examined using an ANOVA approach. The between subjects factor was the number of response categories of a scale (with five levels) while the within subject factor was the meaningful number of magnitude categories (with two levels based on the results of the sorting task across four product attributes, referred to as 3 category and 4 category dimensions). Five (number of points in scale; 3, 4, 5, 7, or 9 points; between subjects) by two (dimensions based on mean number of groups based on sorting task; 3 versus 4 categories; within subjects) by two (replication of dimensions; within subjects) factorial ANOVAs were performed on data based on the number of stimuli captured by 25%, 50%, and 75% of the response categories. These three percentages were chosen to cover the range of possible percentages of response categories.

For the ANOVA on the number of stimuli captured by 25% of the response categories, a significant effect was obtained for dimensions ($F(1,143) = 98.72$; $p < .001$) with higher means (i.e.,

a larger number of stimuli) being found for the 3-category dimensions, thereby providing support for the hypothesis that fewer response categories would be required to capture a set of stimuli. An investigation of simple effects for the number of scale points by dimensions interaction suggested that the effect of dimensions was significant for the nine point scale ($F(1,143) = 57.17$; $p < .001$), the seven point scale ($F(1,143) = 7.58$; $p < .01$), and the five point scale ($F(1,143) = 7.22$; $p < .01$), marginally significant for the four point scale ($F(1,143) = 3.79$; $p < .06$), and nonsignificant for the three point scale ($F(1,143) = 0.88$; $p > .35$) (see Table 1 and Figure 1). Therefore, the hypothesis was supported for 4, 5, 7, and 9 point scales. As suggested earlier, the hypothesis was supported scenarios where the number of response categories was greater than the number of meaningful categories for the 3-category dimension. It also appears that, with an increase in the number of response categories in a scale, the differences for 3-category and 4-category dimensions increases. While no significant difference was obtained for the three point scale, a marginally significant difference was obtained for the 4 point scale (which contains 1 more category than is required for the 3-category dimension) and significant differences were obtained for the 5, 7, and 9 point scales.

Insert Table 1 and Figure 1 about here

A similar ANOVA was performed on the number of stimuli captured by 50 percent of the response categories. A significant effect was obtained for dimensions ($F(1,143) = 108.76$; $p < .001$), with the pattern of means for dimensions providing support for the hypothesis. An investigation of simple effects for the number of scale points by dimensions interaction led to significant effects of dimensions for the nine point scale ($F(1,143) = 30.47$; $p < .001$), the seven point scale ($F(1,143) = 8.01$; $p < .01$), the five point scale ($F(1,143) = 14.39$; $p < .001$) and the four point scale ($F(1,143) = 7.80$; $p < .01$) with a nonsignificant effect for the three point scale ($F(1,143) = 2.65$; $p > .10$). Again, a similar trend is obtained for differences for the 3, 4, 5, 7, and 9 point scales, respectively, providing support for the hypothesis.

A similar ANOVA was performed on the number of stimuli included by 75 percent of the response categories. A significant effect was obtained for the dimensions ($F(1, 143) = 76.16$; $p < .001$). An investigation of simple effects for the number of scale points by dimensions interaction led to non-significant effects of dimensions groups for the nine point scale ($F(1,143) = 2.48$; $p >$

.10), and the seven point scale ($F(1,143) = 2.50$; $p > .10$), and significant effects for the five point scale ($F(1,143) = 12.81$; $p < .001$), the four point scale ($F(1,143) = 22.52$; $p < .001$), and the three point scale ($F(1,143) = 5.28$; $p < .05$). It appears that for the number of stimuli captured by 75% of the response categories, differences between the 4-category and 3-category dimensions were not significant for the 7 and 9 point scales because nearly all the stimuli (mean = 99%) are captured by 75% of the response categories for these two scales, thereby concealing the effect of dimensions.

These results demonstrate that the meaningful number of response categories that respondents use in thinking about a continuum (as determined by the sorting task) influences the number of response categories that are utilized by respondents in completing rating scales. Significant differences in the number of stimuli captured by a certain percentage of response categories were found for the 4, 5, 7, and 9 point scales. In order to assess the extent to which scales with greater number of scale points elicited greater discrimination, the mean number of stimuli captured by the three most frequently used response categories for all four sets of products were compared across the 3, 4, 5, 7, and 9 point scales. The means were 3.99, 3.71, 3.53, 3.13, and 2.97, for the 3, 4, 5, 7, and 9 point scales, respectively, with all pairwise differences being significant at the .01 level. These results suggest that, while the meaningful number of categories for a dimension exerted a downward effect on the number of response categories used in scale response, the number of response categories in a scale exerted an upward effect.

The ratings of each set of products on specific dimensions are presented in Table 2. To demonstrate divergent inferences that could be drawn from different scales as was argued in drawing a distinction between meaningful and maximum discrimination, the following analysis was performed. Pairwise t-tests were performed between pairs of products that were adjacent to each other in terms of the order of their ratings on the 3 point scale for the 3-category dimensions to identify significant or marginally significant differences. Similar tests were also run on the same set of pairs for scales with 7 and 9 response categories. Therefore, the ordering of products according to ratings on the 3 point scale was used as a baseline on the assumption that the number of categories in this scale match the meaningful number of categories For ratings on the 3 point scale of soft drinks, three pairs had significant or marginally significant differences; Sprite and Diet Orange Crush, Diet Orange Crush and Orange Crush, and Diet Pepsi and Pepsi. A similar set of t-tests for

adjacent ratings (using the same ordering) on the 7 point scale suggested differences between these three pairs as well as three additional pairs; Diet Sprite and Sprite, Orange Crush and Diet Coke, and Cherry Coke and Coke Classic. For ratings on 9 point scales, 6 differences were suggested which did not overlap perfectly with the differences based on the 7 point scale. Similarly, for ratings of candy bars, ratings on the three point scale suggested a significant or marginally significant difference for one pair whereas ratings on the seven point scale suggested differences for this pair and an additional pair, while ratings on the 9 point scale suggested differences for this pair and another pair.

Insert Table 2 about here

For ratings of beverages, pairwise differences between adjacent pairs after ordering according to ratings on the 4 point scale (since this was a 4-category dimension) suggested six significant or marginally significant differences. However, ratings on the 7 point scale suggested five such differences while ratings on the 9 point scale suggested six differences, without perfect overlap between any set of pairs that were different. For ratings of snack foods, one significant or marginally significant difference between adjacent pairs was obtained for ratings on the 4 point scale, three for the 7 point scale, and seven for the 9 point scale. Therefore, even though the results suggest that meaningfulness influences scale response in terms of the number of categories that are used by respondents, an effect of the number of scale points in a scale is apparent in that inferences about product ratings may depend on the scale that is used. Several possibilities exist for the different inferences that may be drawn across scales depending on the number of response categories in a scale when compared to the meaningful number of categories for a dimension. Scales with a number of response categories higher than the meaningful number may result in the elicitation of discriminations that may not be meaningful, while scales with too many response categories may result in errors in response that may magnify or even hide meaningful differences.

The rater reliabilities of the scales for each product-dimension were computed according to an analysis of variance procedure suggested by Ebel (1951) and are presented in Table 3.[4] The results do not suggest an increase or a decrease in reliability with an increase in the number of scale points.

Therefore, based on the metric approach to determining the optimal number of response categories in a scale, the various scales with different number of scale points would be considered as being equivalent. However, as discussed earlier, using meaningfulness of the number of categories as a criterion, ratings based in scales with different numbers of response categories may lead to different inferences about product ratings. As suggested earlier, the issue here is not one of reliability of scales, but their validity in eliciting meaningful discriminations, which in turn could be used to make valid inferences about respondents. While the notion of reliability captures the degree of variance in scale response that can be attributed to differences in ratings across products or stimuli, whether these differences are meaningful to respondents is a question that is pertinent to the validity of a scale.[5]

Insert Table 3 about here

Study 3

Study 3 used a procedure that was largely similar to the procedure in Study 2 in an attempt to replicate the results of Study 2. The number of response categories of scales were manipulated using 4 levels, 3, 4, 5, and 7 categories. Also, in contrast to Study 2, the number of response categories of scales used for 3 versus 4 category dimensions was manipulated within subjects. One group of 34 subjects filled out the 3-category dimensions using a 3 point scale and the 4-category dimensions using a 4 point scale while another group of 29 subjects filled out the 3-category dimensions using a 4 point scale and the 4-category dimensions using a 3 point scale. A third group of 24 subjects filled out the 3-category dimensions using a 5 point scale and the 4-category dimensions using a 7 point scale while a fourth group of 24 subjects filled out the 3-category dimensions using a 7 point scale and the 4-category dimensions using a 5 point scale. Subjects rated sets of stimuli for four product-dimensions. The same set of product-dimension combinations were used as in Study 2 except that chocolate flavor of candy bars (which had a mean number of sorted categories of 3.50) was replaced with caramel flavor of candy bars (which had a mean number of sorted categories of 2.75). Therefore, based on the sorting task, two product-dimensions had approximately 3 categories and two had approximately 4 categories.

Results.

ANOVAs were run on each set of two groups using a 2 (match between meaningful number of categories and scale points; 3-3/4-4 versus 3-4/4-3 (or 3-5/4-7 versus 3-7/4-5) where the first digit in each pair refers to the 3 versus 4 category dimension while the second digit refers to the number of scale points of rating scales) by 2 (3-category versus 4-category dimensions) by 2 (replication of attributes) factorial design. For the ANOVA on 3 versus 4 point scales based on the number of stimuli captured by 25% of the response categories, the effect of dimensions was significant for the three point scale ($F(1,116) = 7.81$; $p < .01$), and the four point scale ($F(1,116) = 19.87$; $p < .001$) (see Table 1). For the ANOVA on 5 versus 7 point scales based on the number of stimuli captured by 25% of the response categories, the effect of dimensions was significant for the five point scale ($F(1,66) = 25.32$; $p < .001$), and the seven point scale ($F(1,66) = 16.35$; $p < .001$) (see Table 1). The pattern of means and the significant effect of dimensions provides support for the hypothesis. Similar ANOVAs were run for the number of stimuli captured by 50% of the response categories. Significant effects of dimensions were found for the three point scale ($F(1,115) = 8.40$; $p < .01$), and the four point scale ($F(1,115) = 18.34$; $p < .001$) (see Table 1). The effect of dimensions was also significant for the five point scale ($F(1,66) = 23.78$; $p < .001$), and the seven point scale ($F(1,66) = 6.86$; $p < .05$) (see Table 1). Again, the pattern of means provide support for the hypothesis. For the number of stimuli required captured by 75% of the response categories, the effect of dimensions was significant for the three point scale ($F(1,114) = 6.70$; $p < .05$), and the four point scale ($F(1,114) = 8.00$; $p < .01$) (see Table 1). The effect of dimensions was also significant for the five point scale ($F(1,79) = 23.42$; $p < .001$), but not for the seven point scale ($F(1,79) = 1.15$; $p > .25$) (see Table 1). These results replicate the results of Study 2, providing further support for the hypothesis. One difference in the results between Study 2 and Study 3 was that significant differences between 3 and 4 category dimensions were found for the 3 point scale in Study 3, in contrast to Study 2. The means across the two studies for each condition were also similar with a moderate level of variation. The rater reliabilities for each product-dimension were computed and are presented in Table 3, with the results suggesting neither an increase nor a decrease in reliability with an increase in the number of scale points in a scale.

GENERAL DISCUSSION

This paper suggests the importance of using the number of categories that are meaningful to respondents in thinking about a continuum in designing response scales. It is argued that the measurement of scale responses based on the meaningful number of response categories may be more predictive of variables of importance such as choice and judgment. While scales eliciting a maximum level of discrimination may be more reliable than scales eliciting a meaningful level of discrimination, it is argued that the latter may be more valid in measuring responses that reflect the discriminations that are meaningful to respondents. Drawing from research in psychology on breadth of categorization, a sorting task was used as an independent, open-ended method of understanding the number of response categories that respondents use in thinking about a specific dimension. Using the results of this task as a basis, scale responses to ratings of stimuli along specific dimensions which differed in terms of the number of meaningful categories were collected in a second study using scales with 3, 4, 5, 7, and 9 categories. The results demonstrate the effect of the meaningful number of categories for a dimension on scale response such that a dimension with fewer categories led to the use of fewer scale points to rate stimuli. The results were replicated in a third study. Therefore, the effect of meaningful number of categories on scale response was demonstrated. These results are interesting since they suggest that analyses of scale responses may provide diagnostic information about the meaningful categories for specific dimensions.

It appears that the meaningful number of response categories used by respondents to think about a continuum (as determined using a sorting task) influences the number of response categories used by respondents in rating scales. This result is important in several ways. It brings out the importance of understanding the number of response categories that are meaningful to respondents, and therefore, the immediate need to develop methods at the pretesting stage to evaluate the nature of information being collected using scales with different number of response categories, as suggested by Cox (1980). It also brings out the importance of understanding a new factor (i.e., meaningfulness of response categories) in determining the optimal number of response categories to use in a scale. In addition to limits to human ability to discriminate and researchers' need for information, a third factor that should be considered is the meaningfulness of the number of response categories. The level of discrimination that is elicited for purposes of the researcher should be

tempered by the meaningfulness of such discrimination. Another important point to note is that analyses of scale responses could provide insight into the issue of meaningfulness. Using data analysis aimed at assessing the frequency of usage of scale points, it appears that the effect of meaningfulness may be detected using scale responses.

Several lines of future research are suggested based on the results of this study. First, the relationship between the number of meaningful categories and scale responses needs to be examined. Such analyses of scale responses may provide a basis for developing techniques to extract data that is at a level of discrimination that is meaningful to respondents from scale responses. Using frequency of utilization of scale responses in combination with ratings of stimuli, information could be extracted about meaningful differences between rated stimuli. Further, the relationship between meaningful discrimination, maximum discrimination, and response error should be studied in order to enhance the validity of inferences made from product ratings. Second, additional methods are needed to understand the number of categories that are meaningful to respondents on specific product attributes. While the sorting task appears to be a valid means of assessing meaningfulness of responses, other open-ended methods should be developed and used as a basis in in scale development. Open-ended methods could also provide insights into the nature of the categories used by respondents to think about an attribute continuum, which could be useful in developing category descriptors. Further, research should focus on the relationship between the number of meaningful categories and properties of attributes (such as attribute importance or the concreteness-abstractness of attributes). Such research could provide guidelines for the number of categories that may be meaningful for different types of attributes. In conclusion, the meaningfulness of categories used by respondents in thinking about an attribute continuum is an important factor that should be considered in scale development as well as in the analyses of scale responses.

Footnotes

[1] The distinction between meaningful and maximum discrimination on specific attributes has implications for the prediction of overall judgments based on the combination of information across attributes. Consider a product with 'n' attributes measured with 'x' response categories in one case and 'y' response categories in another, where x and y refer to meaningful versus maximum number of response categories, respectively. In combining information across n attributes, the possible number of levels of an overall judgment, assuming equal weightage for all n attributes, is x^n and y^n, respectively, for the two cases. For a product with two attributes and two meaningful levels of each attribute, where attribute ratings are measured with 7 point scales, the possible number of levels of overall judgments are 4 and 49, respectively. Therefore, differences between meaningful and maximum discrimination at the attribute level grow multiplicatively with the number of attributes at the overall judgment level.

[2] A sample calculation of the number of stimuli captured by 25% of the response categories of a scale is shown below. Using a 5 point scale where the number of stimuli captured by the most frequently used category, the second most frequently used category, and so on are 5, 4, 3, 0, and 0, respectively, the number of stimuli captured by 25% of the response categories was 6 (i.e., 5 + [(.25-.20)/(.40-.20)]x4, the number of stimuli captured by 20% of the response categories (i.e., the most frequently used category) added to the number of stimuli captured by the next 5% of response categories (i.e., 1/4th of the second most frequently used category)).

[3] It should be noted that non-usage or low usage of a category may also occur if one or more stimuli are clearly rated higher than the others. For example, if respondents consider one brand to be a 7 and the next highest brand to be a 5 and, therefore, do not use the response category that is between the two (i.e., a 6), this would appear as non-usage by the present method. In other words, the sorting task only provides the number of categories that are meaningful to respondents but not the magnitude differnces between these categories. However, this possibility can be examined by looking at the means for various stimuli in conjunction with the frequency of usage of each specific response category. If stimuli are uniformly distributed across a continuum, then such an explanation for non-usage of scale points may not hold.

[4] In line with Cox's (1980) recommendation for the use of rater reliability for stimulus-centred scales, reliability was computed as the ratio of the variance between stimuli to the sum of variance between stimuli and residual variance (see Ebel, 1951).

[5] A sample computation of the information transmitted by 3, 5, and 7 point scales for ratings of caffeine content of soft drinks was performed according to the procedure suggested by Garner and Hake (1951) and found to be 0.59, 0.62, and 0.70, for the 3, 5, and 7 point scales, respectively. Therefore, based on this criterion, the 7 point scale would be considered the most appropriate of the three scales considered.

References

Banks, W.P. (1977), "Encoding and Processing of Symbolic Information in Comparative Judgments," in G.H. Bower (ed.), The Psychology of Learning and Motivation, New York: Academic Press, Vol. 11, 101-159.

Bendig, A. W. (1953), "The Reliability of Self-Ratings as Function of the Number of Categories on the Scale," The Journal of Applied Psychology, 37(1), 38-41.

Bendig, A. W. (1954), "Reliability and the Number of Rating Scale Categories," The Journal of Applied Psychology, 38(1), 38-40.

Block, Jack, Buss, David M., Block, Jeanne H., & Gjerde, Per F. (1981). The Cognitive Style of Breadth of Categorization: Longitudinal Consistency of Personality Correlates. Journal of Personality and Social Psychology, 40 (4), 770-779.

Champney, H., and Helen Marshall (1939), "Rater's Minimal Discrimination as a Criterion for Determining the Optimal Refinement of a Rating Scale," Journal of Applied Psychology, 23, 323-331.

Churchill, Gilbert A., and J. Paul Peter (1984), "Research Design Effects on the Reliability of Rating Scales: A Meta-Analysis," Journal of Marketing Research, 21, 360-75.

Cox, Eli (1980), "The Optimal Number of Response Alternatives in a Scale: A Review," Journal of Marketing Research, 17, 407-422.

Ebel, Robert L. (1951), "Estimation of the Reliability of Ratings," Psychometrika, 16, 407-24.

Gardner, R.W., & Schoen, R.A. (1962). Differentiation and Abstraction in Concept Formation. Psychological Monographs, 76 (41, Whole No. 560).

Garner, W. R. (1960), "Rating Scales, Discriminability, and Information Transmission," The Psychological Review, 67(6), 343-352.

Garner, W.R., and H.W. Hake (1951), "The Amount of Information in Absolute Judgments," Psychological Review, 58, 446-459.

Green, Paul E. and Vithala R. Rao (1970), "Rating Scales and Information Recovery-How Many Scales and Response Categories to Use?," Journal of Marketing, 34, 33-39.

Hinrichs, James V., and Laura R. Novick (1982), "Memory for Numbers: Nominal Vs. Magnitude Information," Memory and Cognition, 10 (5), 479-486.

Holyoak, K.J. and W.A. Mah (1982), "Cognitive Reference Points in Judgments of Symbolic Magnitude," Cognitive Psychology, Vol. 14, 328-352.

Jacoby, Jacob, and Michael S. Matell (1971), "Three-Point Likert Scales Are Good Enough," Journal of Marketing Research, 8, 495-500.

Johnson, Michael D. and Claes Fornell (1987), "The Nature and Methodological Implications of the Cognitive Representation of Products," Journal of Consumer Research, Vol. 14 (September), 214-228.

Johnson, Michael D., Donald R. Lehmann, and Daniel R. Horne (1990), "The Effects of Fatigue on Judgments of Interproduct Similarity," International Journal of Research in Marketing, 7, 35-43.

Komurita, S. S., and William K. Graham (1965), "Number of Scale Points and the Reliability of Scale," Educational and Psychological Measurement, 25(4), 987-995.

Parducci, Allen (1965), "Category Judgment: A Range Frequency Model," Psychological Review, 72(6), 407-418.

Ramsay, J.O. (1973), "The Effect of Number of Categories in Rating Scales on Precision of Estimation of Scale Values", Psychometrika, 38 (4), 513-532.

Sudman, Seymour, and Norman M. Bradburn (1982), Asking Questions, San Francisco: Jossey-Bass.

Symonds, P. M. C. (1924), "On the Loss of Reliability in Ratings Due to Coarseness of the Scale," Journal of Experimental Psychology, 7, 456-461.

Tversky, Amos (1977), "Features of Similarity," Psychological Review, 84 (4), 327-352.

Viswanathan, Madhubalan, & Childers, Terry (1992). The Encoding and Utilization of Magnitudes along Product Attributes: An Investigation Using Numerical and Verbal Information. Unpublished Manuscript.

Wyatt, Randall C., and Lawrence S. Meyers (1987), "Psychometric Properties of Four 5-Point Likert-Type Response Scales," Educational and Psychological Measurement, 47, 27-35.

TABLE 1

RESULTS FOR STUDY 2 AND STUDY 3

NO. OF RESPONSE CATEGORIES	TYPE OF DIMENSION					
	3-CATEGORY DIMENSIONS			4-CATEGORY DIMENSIONS		
	PERCENTAGE OF RESPONSE CATEGORIES					
	25%	50%	75%	25%	50%	75%
Study 2						
3	4.69	8.31	10.76	4.38	7.84	10.38
4	5.77	9.42	11.52	5.12	8.62	10.73
5	6.46	10.13	11.71	5.53	9.00	11.10
7	7.28	10.55	11.86	6.33	9.71	11.58
9	8.20	11.19	11.88	5.76	9.66	11.63
Study 3						
3	5.20	8.64	10.82	4.47	7.89	10.40
4	6.54	9.81	11.31	5.38	8.71	10.85
5	7.14	10.40	11.71	5.47	9.07	11.13
7	8.21	10.99	11.80	6.87	10.27	11.67

TABLE 2

RATINGS FOR STUDY 2

PRODUCT-DIMENSION	NUMBER OF RESPONSE CATEGORIES				
	3	4	5	7	9
CAFFEINE CONTENT OF SOFT DRINKS					
1. Diet Seven-Up	1.17	1.24	1.50	1.79	1.69
2. Seven-Up	1.21	1.30	1.83	1.96	2.13
3. Diet Sprite	1.33	1.47	1.68	2.25	2.22
4. Sprite	1.47	1.70	2.03	2.54	2.63
5. Diet Orange Crush	1.83	2.14	2.32	3.18	4.12
6. Orange Crush	2.10	2.43	2.69	3.68	4.61
7. Diet Coke	2.17	2.93	3.57	5.18	6.84
8. Diet Pepsi	2.17	2.93	3.46	5.25	6.72
9. Pepsi	2.87	3.60	4.50	5.89	7.66
10. New Coke	2.87	3.73	4.14	5.82	7.81
11. Cherry Coke	2.93	3.53	4.50	5.82	7.69
12. Coke Classic	2.97	3.67	4.57	6.07	7.78
CHOCOLATY FLAVOR OF CANDY BARS					
1. M & M Peanut	1.87	2.73	3.48	4.30	5.69
2. Twix Caramel	1.87	2.50	3.32	4.04	5.03
3. Mars Bar	1.93	2.55	3.32	4.07	5.26
4. Kit Kat	2.00	2.60	3.54	4.21	5.56
5. Reese's Peanut B Cs	2.03	2.47	3.54	4.32	5.69
6. Three Musketeers	2.30	3.10	3.39	4.61	5.87
7. Milky Way	2.30	3.17	3.82	4.68	6.44
8. Snickers	2.43	3.03	3.79	4.93	6.88
9. M & M Plain	2.63	3.60	4.14	5.89	7.16
10. Nestle's Crunch	2.73	3.48	4.07	6.21	7.22
11. Hershey's Almond	2.76	3.47	4.43	6.00	7.47
12. Hershey's Plain	2.97	3.97	4.71	6.81	8.44
SWEETNESS OF BEVERAGES					
1. Bottled Water	1.00	1.00	1.10	1.04	1.13
2. Milk	1.03	1.23	1.52	1.43	2.12
3. Club Soda	1.13	1.53	1.71	1.79	2.88
4. Iced Coffee	1.38	1.62	1.85	2.04	2.09
5. Iced Tea	1.43	1.73	1.83	2.57	3.78
6. Diet Soft Drink	1.93	2.33	3.03	4.39	4.87
7. Fruit Juice	2.17	2.83	3.45	4.04	5.03
8. Chocolate Milk	2.37	2.90	3.48	4.79	5.56
9. Lemonade	2.00	2.97	3.66	4.14	5.44
10. Milk Shake	2.57	3.10	3.90	5.14	6.66
11. Ice-cream Soda	2.63	3.10	4.00	5.00	6.25
12. Soft Drink	2.63	3.40	3.93	5.57	6.87

TABLE 2- CONTINUED

RATINGS FOR STUDY 2

PRODUCT-DIMENSION	NUMBER OF RESPONSE CATEGORIES				
	3	4	5	7	9
HOW GOOD A SNACK A SNACK FOOD IS					
1. Brownie	1.83	2.03	2.93	3.93	5.19
2. Potato Chips	2.10	2.17	3.17	3.61	6.41
3. Ice-cream	2.03	2.20	3.10	4.04	5.44
4. Candy bar	2.10	2.27	3.00	3.75	5.66
5. Cookie	2.20	2.27	3.28	4.00	6.25
6. Yogurt	1.87	2.67	3.31	3.89	4.63
7. Nacho Chips	2.47	2.70	3.10	5.21	6.38
8. Cheese	2.20	2.73	2.90	4.21	4.81
9. Crackers	2.00	2.80	3.14	4.29	5.41
10. Apple	2.47	3.30	3.79	5.04	5.72
11. Popcorn	2.60	3.47	4.21	5.96	6.97
12. Grapes	2.67	3.47	4.38	5.36	6.22

TABLE 3

RELIABILITIES FOR STUDY 2 & STUDY 3

PRODUCT-DIMENSION	NUMBER OF RESPONSE CATEGORIES				
	3	4	5	7	9
STUDY 2					
1.	0.98	0.98	0.98	0.99	0.99
2.	0.93	0.93	0.88	0.96	0.94
3.	0.97	0.96	0.97	0.98	0.97
4.	0.72	0.80	0.73	0.76	0.64
STUDY 3					
1.	0.98	0.98	0.98	0.96	
2.	0.98	0.98	0.99	0.98	
3.	0.95	0.97	0.96	0.97	
4.	0.64	0.74	0.64	0.67	

Note. Product-dimensions 1, 2, 3, & 4 are caffeine content of soft drinks, chocolaty flavor of candy bars, sweetness of beverages, and how good a snack a snack food is, respectively, in Study 2. Product-dimensions 1, 2, 3, & 4 are caffeine content of soft drinks, caramel flavor of candy bars, sweetness of beverages, and how good a snack a snack food is, respectively, in Study 3.

FIGURE 1A

NUMBER OF STIMULI CAPTURED BY 25 % OF RESPONSE CATEGORIES

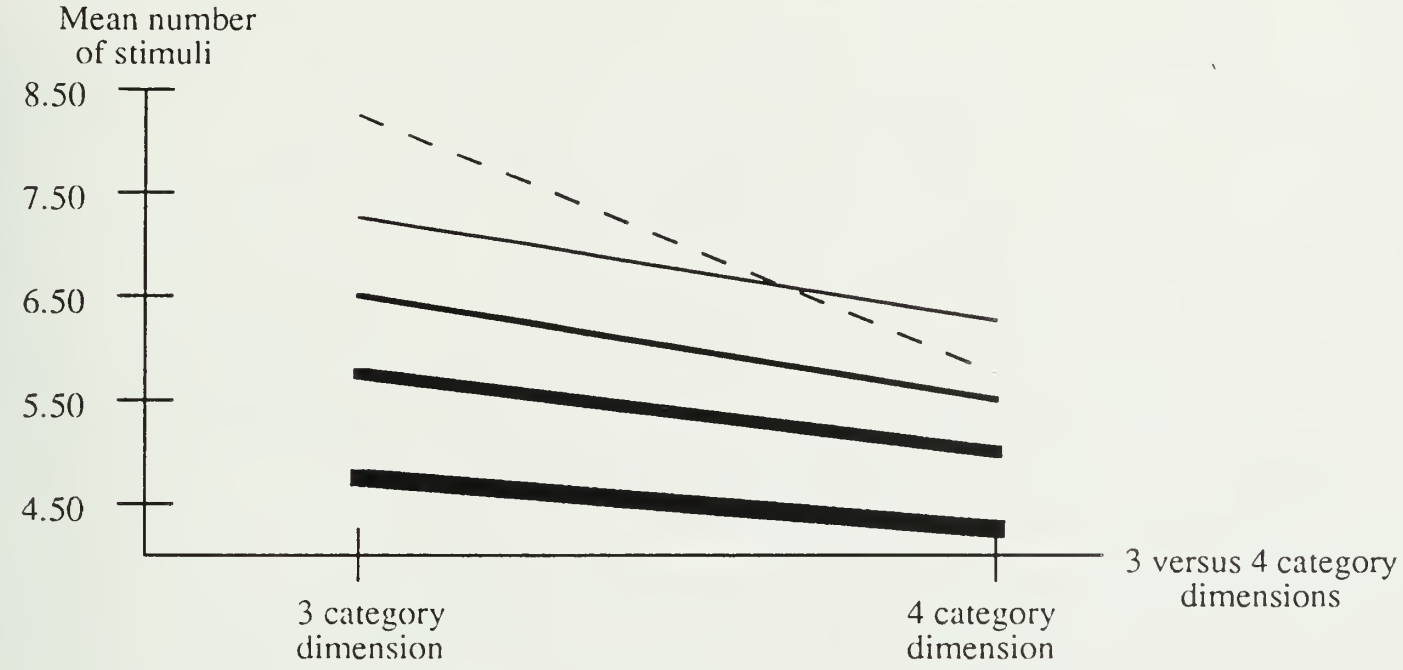
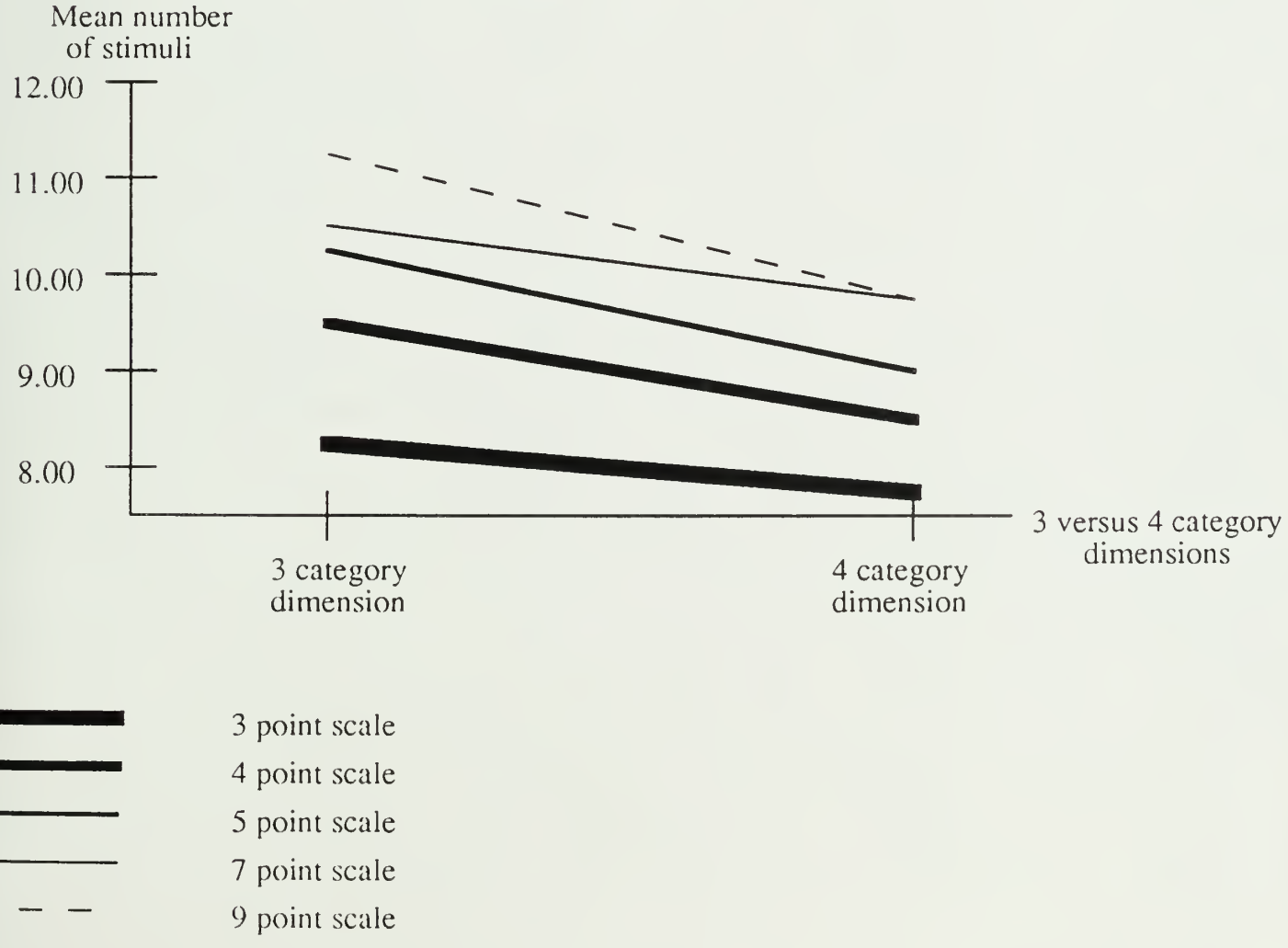


FIGURE 1B

NUMBER OF STIMULI CAPTURED BY 50 % OF RESPONSE CATEGORIES



UNIVERSITY OF ILLINOIS-URBANA
3 0112 060295869